Molecular recognition of DNA by ligands: Roughness and complexity of the free energy profile

Wenwei Zheng,[1,a)] Attilio Vittorio Vargiu,[2,a)] Mary A. Rohrdanz,[1] Paolo Carloni,[3] and Cecilia Clementi[1,b)]
[1]*Department of Chemistry, Rice University, Houston, Texas 77005, USA*
[2]*Department of Physics, University of Cagliari, S.P. Monserrato-Sestu, 09042 Cagliari, Italy*
[3]*German Research School for Simulation Science, GmbH 52425 Jülich, Germany and Research Center and RWTH-Aachen University, Jülich, Germany*

(Received 2 July 2013; accepted 19 September 2013; published online 11 October 2013; corrected 22 November 2013)

Understanding the molecular mechanism by which probes and chemotherapeutic agents bind to nucleic acids is a fundamental issue in modern drug design. From a computational perspective, valuable insights are gained by the estimation of free energy landscapes as a function of some collective variables (CVs), which are associated with the molecular recognition event. Unfortunately the choice of CVs is highly non-trivial because of DNA's high flexibility and the presence of multiple association-dissociation events at different locations and/or sliding within the grooves. Here we have applied a modified version of Locally-Scaled Diffusion Map (LSDMap), a nonlinear dimensionality reduction technique for decoupling multiple-timescale dynamics in macromolecular systems, to a metadynamics-based free energy landscape calculated using a set of intuitive CVs. We investigated the binding of the organic drug anthramycin to a DNA 14-mer duplex. By performing an extensive set of metadynamics simulations, we observed sliding of anthramycin along the full-length DNA minor groove, as well as several detachments from multiple sites, including the one identified by X-ray crystallography. As in the case of equilibrium processes, the LSDMap analysis is able to extract the most relevant collective motions, which are associated with the slow processes within the system, i.e., ligand diffusion along the minor groove and dissociation from it. Thus, LSDMap in combination with metadynamics (and possibly every equivalent method) emerges as a powerful method to describe the energetics of ligand binding to DNA without resorting to intuitive *ad hoc* reaction coordinates.
 [http://dx.doi.org/10.1063/1.4824106]

I. INTRODUCTION

Processes associated with DNA are key targets of intervention against a variety of diseases, cancer being the best known example.[1] Development of effective drugs targeting DNA (in its many forms, e.g., duplexes,[2,3] G-quadruplexes,[4] or more complex structures within the genome[5]) has often been based on experiments on ligand- and protein-DNA complexes, from structural studies to the measurements of kinetic and thermodynamic data.[1]

On the other hand, computer-aided strategies have faced major challenges in addressing the complexity of nucleic acids structures and dynamics.[6,7] Even in the "simple" case of a DNA double-stranded helix, DNA's extreme flexibility (which plays a key role in the molecular recognition event) poses a serious sampling problem. Moreover, recognition by ligands occurs generally via multiple association-dissociation processes between different sites and/or slidings within the grooves.[2,8–10] Dissociation regulates the residence time of ligands in different sites,[9,11–15] and sliding is important to optimize efficacy and selectivity.[7,16] In particular, since association and sliding within the groove are likely to feature lower barriers with respect to dissociation[12–17] (even when the reaction cannot be described via a simple pseudo-first-order kinetics[18]) the free energy profile associated with the latter process is crucial for tuning the dynamic strength of the drug molecule (i.e., the maximum force the complex can resist before dissociation[14,19,20]) and thus the affinity.

Due to the long timescales of these processes, as compared to typical simulation times, enhanced sampling algorithms (see, e.g., Refs. 21–24) are the methods of choice to investigate DNA molecular recognition by small ligands. Most such techniques bias the simulation along a pre-determined set of collective variables (CVs), whose variation is thought to describe the process under investigation. Then, the free energy of the process is calculated as a function of these CVs using a variety of methods, such as umbrella sampling,[25] adaptive biased force,[26] and metadynamics, among others.[6,24,27–29]

The most straightforward CVs are derived from physical or chemical intuition, and serve as approximate variables to gauge the progress of a reaction. The reliability of these intuitive CVs may be assessed by a variety of methods, including the isocommitor surface,[30] genetic neural network algorithm,[31] and Bayesian analysis methods.[32] Alternatively, techniques such as the string methods[33,34] and milestoning[35] have been used to identify reaction pathways, which can be thought of as CVs. However, for all these methods, some

a)W. Zheng and A. V. Vargiu contributed equally to this work.
b)Electronic mail: cecilia@rice.edu

initial choices of the CVs and/or definitions of the reactant and product states are required.

In order to avoid any *a priori* knowledge and intuition about the system when defining a set of CVs, one can extract them from molecular simulation data by using dimensionality reduction methods. These include principle component analysis[36] and its nonlinear variants,[37] local linear embedding,[38] and Isomap.[39,40] Unfortunately, these algorithms encounter difficulties because of the inherent noisiness in biomolecular simulation data.[41] Recently, several new dimensionality reduction algorithms have been introduced and tested against noisy macromolecular processes. The Sketch-map method[42] provides a sketch of the high-dimensional landscape. The Diffusion Map method,[43] and its improved version, Locally Scaled Diffusion Map (LSDMap)[41,44] are able to decouple motions with different timescales into a set of reaction coordinates, named diffusion coordinates (DCs). For systems with a separation of timescales, the first few DCs are sufficient to characterize the slow processes of the system. Reaction rates computed along the 1st DC are in remarkable agreement with the rates measured directly from simulation data,[41,44,45] demonstrating the effectiveness of the LSDMap approach in estimating the barrier heights, and transition rates in biomolecular systems.

It is worth mentioning that instead of using collective variables and obtaining free energy projections on these variables, one could use an alternative technique, namely, cut-based free energy profiles (cFEP),[46,47] to analyze the free energy surface using the partition function of a given region as the progress variable. The cFEP method constructs an equilibrium kinetic network[48] on predefined clusters obtained by criteria such as root mean square deviation (RMSD),[49] secondary structure sequence,[50] and determines the barrier height between different states in the network with methods based on isocommitor surfaces or mean first passage time.[50] The cFEP technique produces a one-dimensional free energy profile between each pair of basins that best preserves the barrier height between the two states and avoids the problems of free energy projections onto specific CVs, which might not cover all motions contributing to the barrier considered. cFEP can also serve as a good check of the quality of the reaction coordinates by comparing the free energy profile obtained by cFEP and the one obtained by projection onto a specific set of reaction coordinates, as shown in Ref. 47.

We recently used both LSDMap and cFEP methods to analyze the folding pathways of a 20-residue three-stranded antiparallel β-sheet peptide called Beta3s.[45] We defined the folding pathways of the system with only the first two DCs and then obtained the physical variables best corresponding to the folding pathways. We found an excellent match between the free energy projected onto these intuitive coordinates and those obtained from the cFEP method. This example illustrates that, at least for the small peptide Beta3s, a set of well-chosen CVs – such as the ones from LSDMap – can significantly reduce the inadequacies of projections of free energy profile into reduced coordinates. One advantage of such a projection is that it allows for an easy interpretation of the pathways and mechanisms between different (meta)stable states.

In the original version of LSDMap, a set of Boltzmann-weighted conformations is required, which does not allow its direct implementation to non-equilibrium simulations biased by intuitive CVs. To address this issue, here we applied an approximate reweighting factor to each configuration generated from a set of metadynamics simulations, allowing treatment of the "reweighted" dataset with LSDMap. In this way the free energy calculated by metadynamics as a function of selected CVs can be projected onto the first few DCs, providing a description of the free energy landscape as a function of unbiased coordinates.

The biological process we investigate here is the molecular recognition of the DNA oligonucleotide d[5′-CAACGTTGGCCAAC-3′]$_2$ by the antibiotic anthramycin in its imino form (hereafter IMI).[51] This system has been previously investigated by some of us with umbrella sampling and metadynamics, which have provided free energy surfaces associated with sliding[16] along the minor groove and dissociation[13] as a function of few (one to three) intuitive CVs. Here, we first extend the number and length of metadynamics simulations, as well as the type of CVs, to increase the reliability of the calculated free energy profiles. We observe sliding of IMI along the whole minor groove length, compared to a length of about three base pairs (bps) spanned in Ref. 16. In addition, we observe detachment from several locations along the minor groove and not only from one as in Ref. 13. The LSDMap analysis prompts us to introduce a new intuitive reaction coordinate which describes the dynamics of the system more precisely than previously used CVs.[13] The free energy surface as a function of the first few DCs extracted from the LSDMap analysis provides a non-empirical view of the system. We identified several distinct diffusion processes at different timescales, the most relevant ones including IMI sliding along the minor groove and its dissociation from the DNA. The LSDMap/metadynamics approach is shown to be a powerful tool to investigate, with an unprecedented level of detail, complicated events such as ligand/DNA dynamic interactions.

II. METHOD

A. All-atom molecular dynamics

The anthramycin · d[5′-CAACGT**TGG**CCAAC-3′]$_2$ (hereafter anthramycin · DNA) system was taken from previously published works.[13,16,51] The AMBER/GAFF force fields[52–54] were used for the parameterization of oligonucleotides and drug. In brief, drug structure was optimized by means of DFT calculations at B3LYP/6-31G(d,p) level, using the Gaussian03 package.[55] Atomic RESP[56] charges were derived using the resp module of AMBER after wavefunction relaxation. See Ref. 51 for details. Potassium ions were modeled with the AMBER-adapted Aqvist potential[57] and the TIP3P model was used for water molecules.[58] Sliding and dissociation of the ligand (anthramycin) were investigated by metadynamics.[27] The CVs used in the implementation of metadynamics in this work are the same as in the previous work.[13] The parameters used in metadynamics are detailed in Sec. I of the supplementary material.[59]

To obtain good statistics and improve the accuracy of the LSDMap reconstruction, 64 independent metadynamics simulations, each of 5 ns in length, were performed (using the GROMACS package[60–62]). Periodic boundary conditions were used, and constant temperature-pressure (T = 300 K, P = 1 atm) dynamics have been performed by the Nosé-Hoover and Andersen-Parrinello-Rahman coupling schemes. Electrostatic interactions were treated using the particle mesh Ewald (PME) algorithm with a real space cutoff of 10 Å, the same as for van der Waals interactions.

Coordinates were saved every 2 ps, and the resulting configurations were filtered by the criterion that the minimum distance between all the heavy atoms of IMI and those of the DNA duplex is smaller than 2.5 nm. This procedure removes less interesting configurations in which the ligand is far away from the DNA duplex. Using this criterium, the data set was reduced to 123 465 conformations.

B. Weighted LSDMap

The LSDMap is based on the kernel

$$K_{ij} = \exp\left(-\frac{\|\boldsymbol{x}_i - \boldsymbol{x}_j\|^2}{2\varepsilon_i \varepsilon_j}\right). \quad (1)$$

In the equation above, $\|\boldsymbol{x}_i - \boldsymbol{x}_j\|$ is the distance between the two configurations $\boldsymbol{x}_i$ and $\boldsymbol{x}_j$. For this system the water molecules within the minor groove of the DNA are essential for a proper description of the ligand-target interactions. Therefore, the RMSD calculation incorporates the DNA, ligand, and those water molecules within the minor groove. The details can be found in Sec. III of the supplementary material.[59] ε_i is the local scale for the configuration $\boldsymbol{x}_i$. This local scale ε_i represents the radius in configuration space around $\boldsymbol{x}_i$ within which the underlying manifold can be approximated by a hyperplane tangent to the manifold, i.e., is approximately linear. The procedure to estimate the local scale around every point in the data set is detailed in previous work.[41] The kernel K_{ij} is related to the "ease" with which $\boldsymbol{x}_i$ can diffuse into $\boldsymbol{x}_j$. A normalized version of this kernel represents the Markov matrix for the dataset of molecular configurations, and the diagonalization of such a matrix yields a set of vectors that serve as DCs.

When the dataset is sampled with a biased statistics, such as metadynamics, the LSDMap algorithm needs to be modified to take into account the bias. Here, we correct for the bias by assigning a weight for each configuration in the dataset, as it has been proposed recently by Ferguson *et al.*[63] In particular, we use a modified version of the algorithm, by defining a symmetric kernel

$$W_{ij} = \sqrt{w_i w_j} K_{ij}, \quad (2)$$

where w_i and w_j are the weights assigned to configurations i and j. The use of a symmetric weighted kernel allows for a much faster and robust eigenvalue decomposition of the corresponding symmetric matrix. This is important when the number of points in the data set is large.

III. RESULTS

A. Landscape of molecular recognition

Fig. 1(a) shows the free energy as a function of the first three DCs. Three narrow low free energy pathways, almost orthogonal to each other, can be seen along the three axes. To identify stable states along the three pathways, it is convenient to project the free energy profiles along each DC, 1st DC (Fig. 1(b)), 2nd DC (inset in Fig. 1(c)), and 3rd DC (inset in Fig. 1(d)). The interesting states along these three coordinates are denoted by uppercase Roman letters, and their representative configurations are shown in Fig. 1(e).

There are two free energy minima, states *A* and *E*, along the 1st DC. State *A* corresponds to the initial configurations in which the ligand binds to the triplet $T_6T_7G_8$, whereas state *E* is associated with binding to the triplet $G_9C_{10}C_{11}$. The three regions *B*, *C*, and *D* along the transition barrier correspond to the ligand sliding along the DNA bases T_7, G_8, G_9, and C_{10}. Typical configurations for these regions (analogous to Fig. 1(e)) are shown in Fig. S1 of the supplementary material.[59] Thus, the 1st DC describes the sliding of the ligand along the central section of the DNA minor groove, and the sliding barrier between state *A* and *E* is approximately 9 kcal/mol (Fig. 1(b)).

Regions *A*, A_1, A_2, and A_3 correspond to four minima along the collective motion characterized by the 2nd DC in Fig. 1(c). States A_1 and A_3 correspond to configurations in which the ligand binds, respectively, to the triplets $G_5T_6T_7$ and $A_2A_3C_4$, with state A_2 in between. This shows that the 2nd DC characterizes the sliding motion along the top section of the DNA. The sliding barrier between state *A* and A_3 is approximately 7.5 kcal/mol (inset in Fig. 1(c)).

Let us define regions E_1, E_2, and E_3 along the 3rd DC. It is clear that this DC corresponds to the ligand sliding along minor groove in the bottom part of the DNA duplex. Regions E_1 and E_3 correspond to configurations in which the ligand interacts with nucleobases C_{10}–A_{12} and A_{12}–C-3′ (the 3′ end of DNA backbone). The sliding barrier between state *E* and E_3 is approximately 15 kcal/mol (inset in Fig. 1(d)). This value is less trustworthy than the barrier heights between other regions because of more limited sampling here.

B. High order DCs

The 4th DC describes an alternative motion of the ligand in the proximity of the bottom part of the duplex (see Fig. 2). Configurations in region E_4 correspond to the detaching of the ligand from the triplet $C_{11}A_{12}A_{13}$ (region E_1), whereas configurations in region E_5 display the ligand re-binding to the duplex at the end of the strands (Fig. S1 of the supplementary material[59]). That is, the 4th DC corresponds to a motion where the ligand partially detaches from the minor groove and then rebinds to 3′ end of DNA.

The 5th DC corresponds to the detachment of anthramycin from the minor groove (see Fig. 2). Region F is located at the negative extreme of the 5th DC, and describes the unbound state of the ligand. Although the lowest free energy path is associated with detachment from the initial

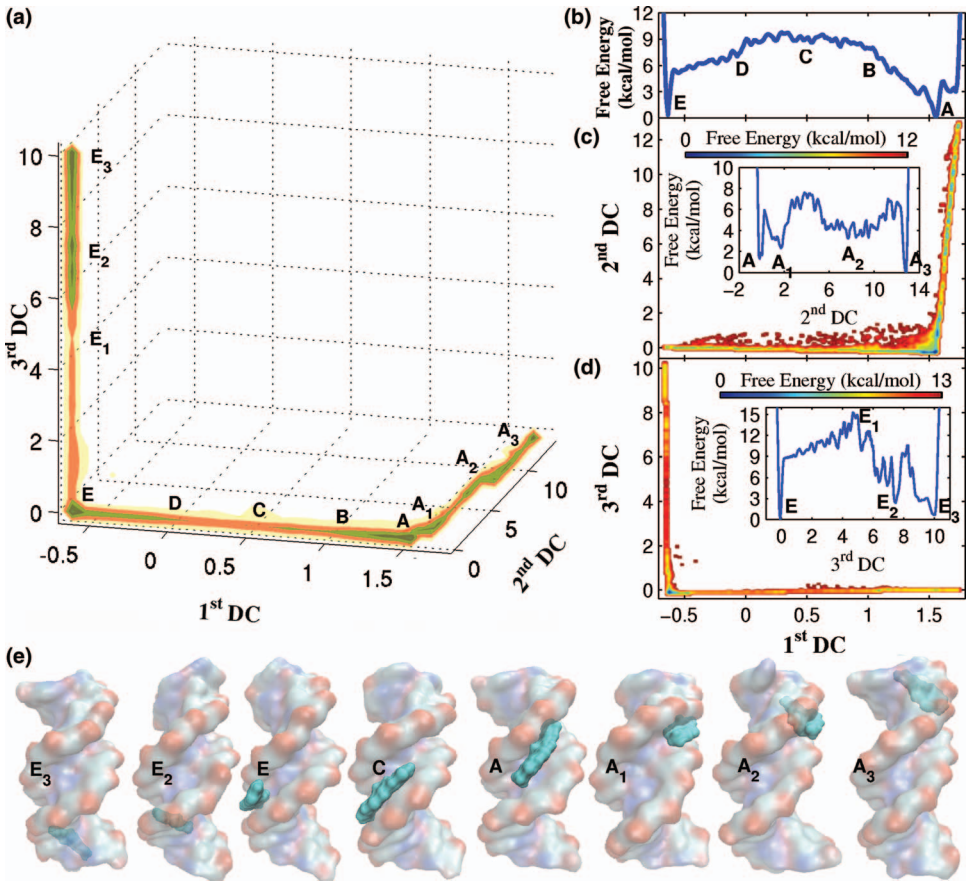

FIG. 1. (a) Free energy projection (in units of kcal/mol) onto the first three DCs. States are as marked. (b) Free energy projection onto the 1st DC. (c) Free energy projection onto the 1st DC and 2nd DC, and 2nd DC (inset). (d) Free energy projection onto the 1st DC and 3rd DC, and 3rd DC (inset). (e) Typical configurations picked in the states marked in the free energy profile.

position A, dissociation can occur from multiple locations along the DNA strands. We will return to this aspect in Sec. III C, after introducing a different physically motivated reaction coordinate. Most of the remaining high-order DCs, which are not shown here, correspond to the same detaching motion of anthramycin from the minor groove as represented by the 5th DC, except that the extrema of these DCs correspond to different configurations. The detaching barrier along 5th DC can be estimated from the free energy profile projected onto this direction. Its value is about 15 kcal/mol, in fair agreement to the value of 12 kcal/mol in Ref. 13. The difference could be due mainly to the following reasons: (a) the barrier calculated here groups detaching motions from several different binding sites along the minor groove, compared to that from a single binding site in Ref. 13; (b) poor sampling becomes an issue when the ligand is far from the initial binding site, therefore the resulting free energy profile must be rather approximate in that region.

C. CV based on H-bonds patterns

Exploiting the insights offered by the LSDMap analysis, we introduced a new intuitive reaction coordinate $\bar{I}$, which approximates the index of the closest DNA bp H-bonding to the ligand, to identify the binding sites on the minor groove. When binding to the DNA duplex, the ligand can form H-bonds with several DNA bps. If for one specific configuration, the ligand forms N_i H-bonds with the ith bp in the DNA duplex, we can define the average bp index

$$\bar{I} = \frac{\sum_{i=1}^{14} N_i \cdot i}{\sum_{i=1}^{14} N_i}. \tag{3}$$

By weighting i by the number of H-bonds formed between the ith DNA bp and the ligand, $\bar{I}$ approximates the index of the closest DNA bp to the ligand in that configuration. If no H-bonds are formed between the ligand and the DNA duplex, $\bar{I}$ is set to zero. Therefore, $\bar{I}$ can also be used to distinguish the binding and detaching configurations between the ligand and the DNA duplex. Here H-bonds are defined with the following cutoffs: $<30^\circ$ for the acceptor-donor-hydrogen angle and <0.35 nm for the distance between donor and acceptor. OH and NH groups are regarded as donors, and O and N are regarded as acceptors.

Fig. 3 shows the value of $\bar{I}$ as a function of the first five DCs. From Fig. 3(a) it is clear that the configurations in which the ligand is bound to the DNA are gathered within two narrow pathways along the first two DCs. $\bar{I}$ varies from 6 to 10 along the 1st DC and from 2 to 6 along the 2nd DC. Another narrow pathway is detected along the 3rd DC shown in Fig. 3(b), where $\bar{I}$ varies from 10 to almost 14 along the 3rd DC. Some fluctuations are obviously present, which are inherent in the definition of $\bar{I}$. Indeed, the ligand can usually

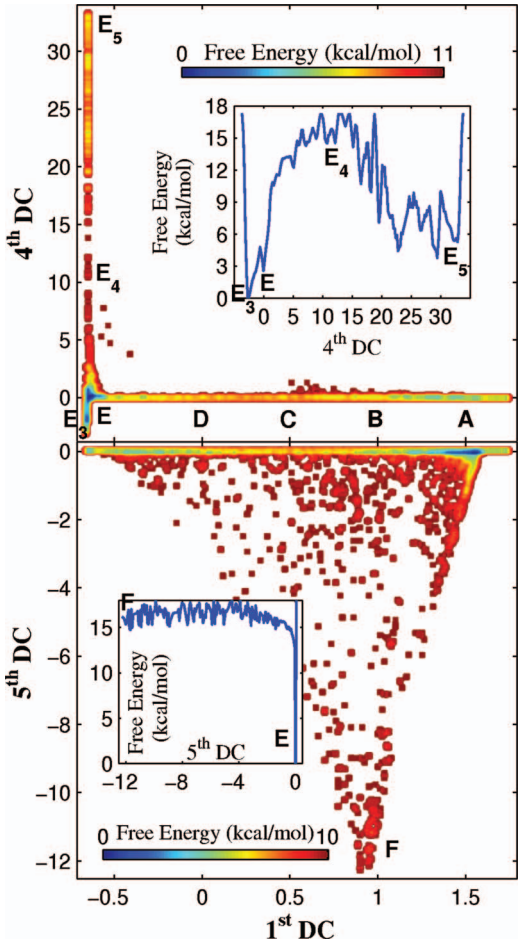

FIG. 2. (Upper panel) Free energy projection (in units of kcal/mol) onto the 1st DC and 4th DC, and 4th DC (inset). States are as marked. (Lower panel) Free energy projection onto the 1st DC and 5th DC (c), and the 5th DC (inset).

form transient H-bonds with two or three consecutive bps. A visual inspection of the configurations without H-bonds between DNA and ligand reveals that the latter can detach from almost any site along the 1st DC, while dissociation is less likely to occur along the 2nd DC and 3rd DC (see Fig. 3). This is consistent with the analysis of the sliding pathways reported above (see Fig. 1), and confirms that detachment from the ends of the duplex is less frequent in our simulations (see also Fig. S2 of the supplementary material[59]). Along the 4th DC and 5th DC (and most of the remaining higher order DCs) there are multiple detaching sites. As expected from the results reported in Sec. III B, the analysis of $\bar{I}$ along the 4th DC confirms that the ligand detaches from the triplets $C_{10}C_{11}A_{12}$, corresponding to state E_4, and then rebinds to the DNA near region E_5. The plot of 1st DC vs. 5th DC (Fig. 3(d)) confirms that anthramycin can detach from every site when sliding along the central part of the minor groove (G_5TTGGC_{10}). However, the projections of $\bar{I}$ onto the 2nd DC-5th DC and 3rd DC-5th DC planes (Figs. 3(e) and 3(f)) show the same behavior as the projections onto the 1st DC-2nd DC and 1st DC-3rd DC, i.e., a low propensity of the ligand to detach from duplex ends. This could be due to the inability of CVs used in metadynamics at introducing a bias strong enough to cause dissociation of anthramycin from those regions of the DNA.

D. Comparison with previous free energy calculations

It is instructive to compare the sliding barriers found here with those shown in Fig. 1 of Ref. 16, reporting the free energy profile associated with the sliding of anthramycin along about 3 DNA bps. Note that the DNA bp numbering scheme we used differs from that in Ref. 16 by two units, so the triplet $T_6T_7G_8$ there corresponds to $T_8T_9G_{10}$ here. Hereafter, we will use our notation also when referring to previously published data. Once this correspondence has been established, it is seen that the sliding among binding sites T_8 (T_6 in Ref. 16), T_9 (T_7), and G_{10} (G_8) corresponds to the transition along states A_1, A, and B. Moreover, the free energy difference and the relative barrier between states A_1 and A, 1.0 and 4.6 kcal/mol, respectively, compare well with the values of 1.5 and 4.0 kcal/mol between states *I V* and *II* in Ref. 16. Also the barrier between states A and B (7.5 kcal/mol) is in fair agreement with that of 5.5 kcal/mol between states *II* and *I*, although in this case the state B (I) does not correspond to a local minimum. These differences may arise from the quality of the sampling and in part from the different reaction coordinates used to estimate the free energy profile.

We next compare our results with the previous metadynamics simulation[13] on this system. In that study the free energy profile associated with ligand unbinding from the triplet $T_8T_9G_{10}$ was generated as a function of two intuitive CVs, namely, the distance d_{CMs} between the centers of mass of the ligand and the DNA tract $d[GTTGG]_2$, and the number of hydrophobic contacts n_{hph} between nonpolar carbons on the ligand and on the bps it covered in the starting structure (Fig. 4). Here we performed a series of 64 independent metadynamics simulations using the same set of CVs as in Ref. 13 (see Sec. I of the supplementary material[59] for details). Our data are in fair agreement with previous results. In particular, the global minimum (labeled I) and the metastable state (*II*) in the upper-left quadrant of Fig. 4 resemble those shown in Fig. 1 of Ref. 13. However, the transition region defined by the two CVs is, as expected, much better sampled here. Moreover, the absence of any wall forcing dissociation from the minor groove allows for the sliding of the ligand along the DNA minor groove. This introduces some differences in other regions of the free energy surface, as compared to previous work. In particular, a new minimum (III) appears, which includes two groups of configurations. The first one involves structures where the ligand is still bound to the DNA, but to some different nucleotide sequence than the initial one. This group is associated with a sliding along the minor groove, up or down with respect to the initial configuration (we only show the configuration sliding up in the lower panel of Fig. 4). The second group involves conformations in which the ligand is partially detached from the DNA, regardless of its position along the groove (we show the ligand detaching from the downside of the DNA duplex in the lower panel of Fig. 4). Thus, the two CVs used in metadynamics cannot distinguish between these configurations, both having large d_{CMs} and near-zero n_{hph}. In contrast to the two intuitive CVs, LSDMap does not take as input any *a priori* knowledge of the system and gives clear separation of the sliding and detaching motions. The relation between the CVs used in metadynamics and the LSDMap

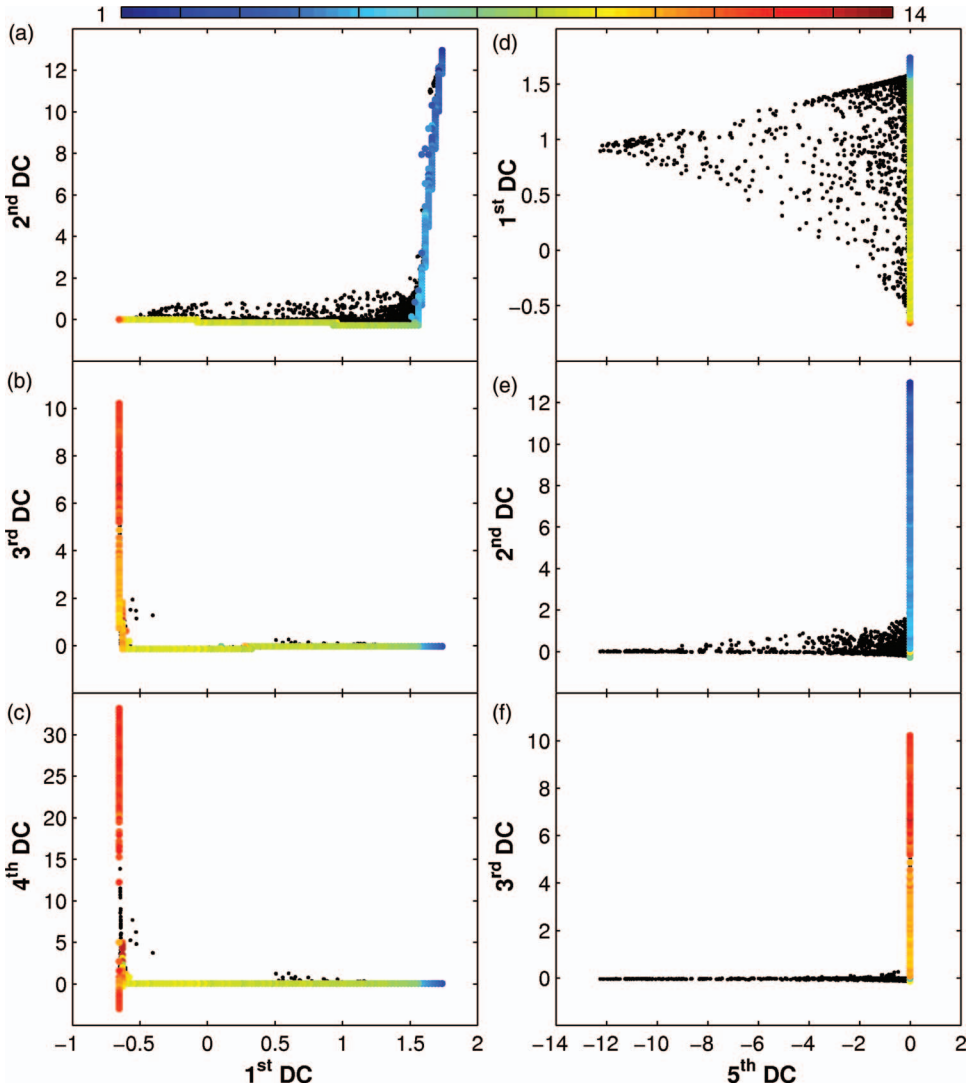

FIG. 3. All the configurations in the data set are projected onto the 1st DC and 2nd DC (a), the 1st DC and 3rd DC (b), the 1st DC and 4th DC (c), the 1st DC and 5th DC (d), the 2nd DC and 5th DC (e), and the 3rd DC and 5th DC (f). The colors indicate the average base pair index $\bar{I}$. The black dots indicate that there are no hydrogen bonds between the ligand and the DNA bases for those configurations. Because the large number of points in the data set introduces a lot of overlaps when $\bar{I}$ is plotted as a function of the DCs, we split the two DCs into 100 × 100 grids and plot the average value of $\bar{I}$ in each bin. The black dots are plotted first so that the colored dots are not covered by the black dots.

coordinates is shown in Fig. S3 of the supplementary material.[59] As expected, we find no clear correlations between the first two DCs and the CVs used in metadynamics.

E. Limitations

Apart from limitations of the force field (see, e.g., Ref. 7), several other factors affect both the quality of the free energy profile and the order of DCs.

First, despite the large number of simulations performed here, the one-dimensional free energy profiles in Fig. 1 are very rough. In particular, that for the 3rd DC displays large fluctuations because of the poorer sampling of the corresponding area of the phase space as compared to that representing initial configuration (state A in Fig. 1(e)). Therefore, the free energy barrier characterized by the 3rd DC is less trustworthy compared to barriers characterized by the 1st DC or 2nd DC, as mentioned in Sec. III.

Second, from an analysis of the order of DCs, one could conclude that the detachment of the ligand from the DNA duplex occurs in a shorter time compared to the sliding of the ligand along the minor groove. However, the free energy barrier is higher for detachment than for sliding, as estimated from the free energy profiles reported here and in previous work.[13] This apparent conflict is very likely a result of unbalanced sampling between configurations of the ligand-DNA complex vs. conformations in which the ligand and DNA are unbound (see Fig. S2 of the supplementary material[59]). The simulations were initiated from structures in which the ligand was bound to the DNA, and were stopped when the ligand dissociated, resulting in many more configurations in the dataset corresponding to the ligand/DNA complex than to the unbound moieties. Though we have unbiased the data when doing LSDMap, the quality of sampling still affected the time scales of the first several slowest motions of the system.

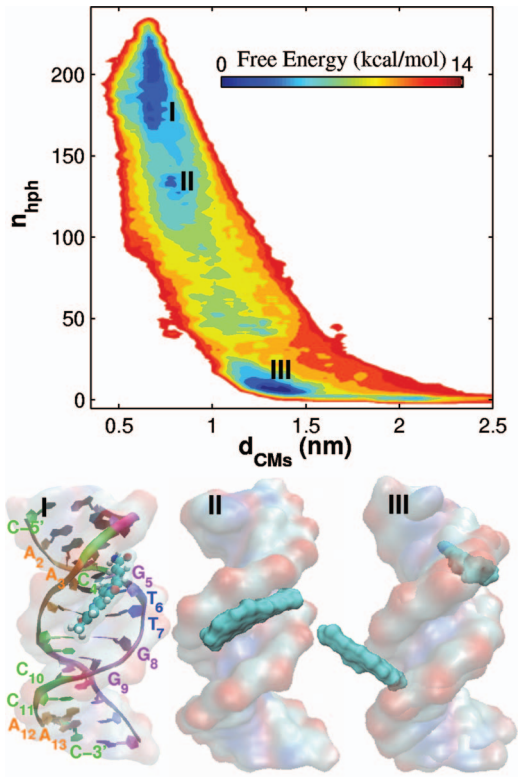

FIG. 4. (Upper panel) Free energy as a function of the two collective variables used in metadynamics, that is, the distance d_{CMs} between the center of mass of the ligand and of the DNA tracts $d[GTTGG]_2$ and the number of hydrophobic contacts n_{hph} between nonpolar carbons on the ligand and on the bases covered by the ligand in the starting structure. (Lower panel) The typical configurations picked in the region as marked in the free energy plot in the upper panel. The two overlapping configurations marked **III** show two possible configurations in region **III**.

Third, for the ease of the interpretation we use only the first few DCs when projecting the free energy landscape, and as a consequence the barrier of the free energy profile might be underestimated due to the neglected contribution of other important motions. This issue is common when using reduced coordinates to represent free energy profiles (see, e.g., Fig. 3 in Ref. 47). However, we believe that this issue is minimal for the system considered here as the DNA and ligand are almost rigid and do not experience severe conformational changes; the fast processes that are neglected in analysis mostly correspond to atomic vibrations in the DNA or ligand. We leave further discussion and more detailed comparison of the performance of LSDMap with respect to other methods, such as, cFEP, to future work.

IV. CONCLUSIONS

In this work, we have investigated the molecular recognition of the oligonucleotide $d[5'\text{-}CAACGTTGGCCAAC\text{-}3']_2$ by anthramycin in its imino form. For this purpose, we performed multiple metadynamics simulations forcing the escape of the ligand from the initial binding position, approximately in the center of the duplex.

The definition of good reaction coordinates is crucial to build a consistent free energy profile. Here, we have used the LSDMap approach[41] for the purpose of identifying a reaction coordinate able to distinguish among relevant conformations associated with ligand-DNA interactions. The most relevant modes found by LSDMap turn out to be associated with the slow processes of the system. In particular, the first three DCs correspond to the diffusion of ligand within the minor groove and the 4th DC and 5th DC correspond to dissociation from multiple sites along the DNA duplex. The LSDMap method allows us to critically assess the quality of the CVs selected for the metadynamics runs. On the basis of that analysis, a new variable is introduced, which is able to discern among different conformations of the system, and is well coupled to the largest eigenvalue modes from the LSDMap analysis.

We have characterized the sliding of anthramycin over the whole minor groove length. The one-dimensional free energy barriers and the metastable states near the initial configuration are in good agreement with those found by some of us in previous umbrella sampling simulations[16] in which the sliding of the ligand along three DNA bps was sampled. The free energy barrier associated with one-dimensional diffusion is lower than that associated with detachment, consistent with previous studies.[13] These results point to the possibility for a ligand to first bind to the DNA duplex in a non-optimal location, and then slide along the groove to the preferred binding site.

We show that LSDMap can be applied to both equilibrium and biased simulations, and provides a solid background for the building of good reaction coordinates even for complex dynamical systems. The combination of metadynamics and LSDMap can be used to study other macromolecular systems with collective diffusion processes in different time scales.

ACKNOWLEDGMENTS

This work was supported by NSF (CDI-type I Grant Nos. 0835824 and CHE-1152344 to C.C.), and the Welch Foundation (C-1570 to C.C.), as well as the Italian NanoCancer Grant from FVG (to P.C.). The original metadynamics simulations were carried out on the Cybersar CPU-cluster at the University of Cagliari. Simulations and other computations were performed on the following shared resources at Rice University: the Cyberinfrastructure for Computational Research funded by NSF under Grant No. CNS-0821727; the Shared University Grid at Rice University funded by NSF under Grant No. EIA-0216467 and in partnership between Rice University, Sun Microsystems, and Sigma Solutions, Inc.; and BlueBioU supported in part by NIH Award No. NCRR S10RR02950 and an IBM Shared University Research (SUR) Award in partnership with CISCO, Qlogic, and Adaptive Computing.

[1]J. B. Chaires, A. Randazzo, and J. L. Mergny, Biochimie **93**, v (2011).
[2]P. B. Dervan, Bioorg. Med. Chem. **9**, 2215 (2001).
[3]Y. H. Du, J. Huang, X. C. Weng, and X. Zhou, Curr. Med. Chem. **17**, 173 (2010).
[4]D. Monchaud and M. P. Teulade-Fichou, Org. Biomol. Chem. **6**, 627 (2008).
[5]T. C. Jenkins, Curr. Med. Chem. **7**, 99 (2000).
[6]X. Biarnés, S. Bongarzone, A. V. Vargiu, P. Carloni, and P. Ruggerone, J. Comput.-Aided Mol. Des. **25**, 395 (2011).
[7]A. Pérez, F. J. Luque, and M. Orozco, Acc. Chem. Res. **45**, 196 (2012).

[8]M. Zewail-Foote and L. H. Hurley, J. Am. Chem. Soc. **123**, 6485 (2001).
[9]S. Y. Breusegem, F. G. Loontiens, P. Regenfuss, and R. M. Clegg, Methods Enzymol. **340**, 212 (2001).
[10]R. Baliga and D. M. Crothers, Proc. Natl. Acad. Sci. U.S.A. **97**, 7814 (2000).
[11]R. Baliga, E. E. Baird, D. M. Herman, C. Melander, P. B. Dervan, and D. M. Crothers, Biochemistry **40**, 3 (2001).
[12]R. Baliga and D. M. Crothers, J. Am. Chem. Soc. **122**, 11751 (2000).
[13]A. V. Vargiu, P. Ruggerone, A. Magistrato, and P. Carloni, Nucleic Acids Res. **36**, 5910 (2008).
[14]G. G. Holman, M. Zewail-Foote, A. R. Smith, K. A. Johnson, and B. L. Iverson, Nat. Chem. **3**, 875 (2011).
[15]L. Wolf, Y. Gao, and R. Georgiadis, J. Am. Chem. Soc. **129**, 10503 (2007).
[16]A. V. Vargiu, P. Ruggerone, A. Magistrato, and P. Carloni, Biophys. J. **94**, 550 (2008).
[17]A. Mukherjee, R. Lavery, B. Bagchi, and J. T. Hynes, J. Am. Chem. Soc. **130**, 9747 (2008).
[18]P. Tummino and R. Copeland, Biochemistry **47**, 5481 (2008).
[19]E. Evans and K. Ritchie, Biophys. J. **72**, 1541 (1997).
[20]A. R. Urbach, Nat. Chem. **3**, 836 (2011).
[21]Y. Okamoto, J. Mol. Graphics Modell. **22**, 425 (2004).
[22]C. Chipot and A. Pohorille, *Free Energy Calculations: Theory and Applications in Chemistry and Biology*, Springer Series in Chemical Physics Vol. 86 (Springer, 2007).
[23]T. Schlick, F1000 Biol Rep **1**, 51 (2009).
[24]V. Leone, F. Marinelli, P. Carloni, and M. Parrinello, Curr. Opin. Struct. Biol. **20**, 148 (2010).
[25]G. M. Torrie and J. P. Valleau, J. Comput. Phys. **23**, 187 (1977).
[26]E. Darve and A. Pohorille, J. Chem. Phys. **115**, 9169 (2001).
[27]A. Laio and M. Parrinello, Proc. Natl. Acad. Sci. U.S.A. **99**, 12562 (2002).
[28]F. Marinelli, F. Pietrucci, A. Laio, and S. Piana, PLoS Comput. Biol. **5**, e1000452 (2009).
[29]M. Ceccarelli, A. V. Vargiu, and P. Ruggerone, J. Phys. Condens. Matter **24**, 104012 (2012).
[30]R. Du, V. S. Pande, A. Y. Grosberg, T. Tanaka, and E. S. Shakhnovich, J. Chem. Phys. **108**, 334 (1998).
[31]A. Ma and A. R. Dinner, J. Phys. Chem. B **109**, 6769 (2005).
[32]R. B. Best and G. Hummer, Proc. Natl. Acad. Sci. U.S.A. **102**, 6732 (2005).
[33]W. E, W. Ren, and E. Vanden-Eijnden, Phys. Rev. B **66**, 052301 (2002).
[34]E. Weinan, W. Ren, and E. Vanden-Eijnden, J. Phys. Chem. B **109**, 6688 (2005).
[35]A. K. Faradjian and R. Elber, J. Chem. Phys. **120**, 10880 (2004).
[36]I. T. Jolliffe, *Principal Components Analysis* (Springer-Verlag, 1986).
[37]Y. Mu, P. H. Nguyen, and G. Stock, Proteins **58**, 45 (2005).
[38]S. T. Roweis and L. K. Saul, Science **290**, 2323 (2000).
[39]J. B. Tenenbaum, V. De Silva, and J. C. Langford, Science **290**, 2319 (2000).
[40]P. Das, M. Moll, H. Stamati, L. E. Kavraki, and C. Clementi, Proc. Natl. Acad. Sci. U.S.A. **103**, 9885 (2006).
[41]M. A. Rohrdanz, W. Zheng, M. Maggioni, and C. Clementi, J. Chem. Phys. **134**, 124116 (2011).
[42]M. Ceriotti, G. Tribello, and M. Parrinello, Proc. Natl. Acad. Sci. U.S.A. **108**, 13023 (2011).
[43]R. R. Coifman, I. G. Kevrekidis, S. Lafon, M. Maggioni, and B. Nadler, Multiscale Model. Simul. **7**, 842 (2008).
[44]W. Zheng, M. A. Rohrdanz, M. Maggioni, and C. Clementi, J. Chem. Phys. **134**, 144109 (2011).
[45]W. Zheng, B. Qi, M. A. Rohrdanz, A. Caflisch, A. R. Dinner, and C. Clementi, J. Phys. Chem. B **115**, 13065 (2011).
[46]S. V. Krivov and M. Karplus, J. Phys. Chem. B **110**, 12689 (2006).
[47]S. Muff and A. Caflisch, Proteins **70**, 1185 (2008).
[48]S. V. Krivov and M. Karplus, J. Chem. Phys. **117**, 10894 (2002).
[49]B. Qi, S. Muff, A. Caflisch, and A. R. Dinner, J. Phys. Chem. B **114**, 6979 (2010).
[50]S. V. Krivov, S. Muff, A. Caflisch, and M. Karplus, J. Phys. Chem. B **112**, 8701 (2008).
[51]A. V. Vargiu, P. Ruggerone, A. Magistrato, and P. Carloni, J. Phys. Chem. B **110**, 24687 (2006).
[52]W. D. Cornell, P. Cieplak, C. I. Bayly, I. R. Gould, K. M. Merz, D. M. Ferguson, D. C. Spellmeyer, T. Fox, J. W. Caldwell, and P. A. Kollman, J. Am. Chem. Soc. **117**, 5179 (1995).
[53]J. Wang, R. M. Wolf, J. W. Caldwell, P. A. Kollman, and D. A. Case, J. Comput. Chem. **25**, 1157 (2004).
[54]D. A. Case, T. E. Cheatham III, T. Darden, H. Gohlke, R. Luo, K. M. Merz, Jr., A. Onufriev, C. Simmerling, B. Wang, and R. J. Woods, J. Comput. Chem. **26**, 1668 (2005).
[55]M. J. Frisch, G. W. Trucks, H. B. Schlegel *et al.*, Gaussian 03, Revision C.02, Gaussian, Inc., Wallingford, CT, 2004.
[56]C. I. Bayly, P. Cieplak, W. Cornell, and P. A. Kollman, J. Phys. Chem. **97**, 10269 (1993).
[57]J. Aaqvist, J. Phys. Chem. **94**, 8021 (1990).
[58]W. L. Jorgensen, J. Am. Chem. Soc. **103**, 335 (1981).
[59]See supplementary material at http://dx.doi.org/10.1063/1.4824106 for more details.
[60]H. Berendsen, D. van der Spoel, and R. Van Drunen, Comput. Phys. Commun. **91**, 43 (1995).
[61]E. Lindahl, B. Hess, and D. van der Spoel, J. Mol. Model. **7**, 306 (2001).
[62]D. van der Spoel, E. Lindahl, B. Hess, G. Groenhof, A. Mark, and H. Berendsen, J. Comput. Chem. **26**, 1701 (2005).
[63]A. L. Ferguson, A. Z. Panagiotopoulos, P. G. Debenedetti, and I. G. Kevrekidis, J. Chem. Phys. **134**, 135103 (2011).